Ridge parameter for g-prior distribution in probit mixed models with collinearity

Meïli Baragatti[1,2,*] and Denys Pommeret[2]

[1] *Ipsogen SA, Luminy Biotech Entreprises, Case 923, Campus de Luminy, 13288 Marseille Cedex 9, France.*

[2] *Institut de Mathématiques de Luminy (IML), CNRS Marseille, case 907, Campus de Luminy, 13288 Marseille Cedex 9, France.*

* *baragatt@iml.univ-mrs.fr, baragattimeili@hotmail.com.*

PREPRINT

Abstract

In the Bayesian variable selection framework, a common prior distribution for the regression coefficients is the g-prior of Zellner [1986]. However, there are two standard cases in which the associated covariance matrix does not exist, and the conventional prior of Zellner can not be used: if the number of observations is lower than the number of variables (large p and small n paradigm), or if some variables are linear combinations of others. In such situations a prior distribution derived from the prior of Zellner can be used, by introducing a ridge parameter. The prior obtained is a flexible and simple adaptation of the g-prior, and can be linked to the work of Gupta and Ibrahim [2007]. In this paper a simple way to choose the associated hyper-parameters is proposed, and a full variable selection method using this prior is developed for probit mixed models. The method is then applied to both simulated and real datasets in which some variables are linear combinations of others.

Keywords: Bayesian variable selection, Zellner prior, ridge parameter, probit mixed regression model, grouping technique (or blocking technique), Metropolis-within-Gibbs algorithm.

1 Introduction

We consider the problem of variable selection in a probit mixed model with Y a n-vector of responses, given a set of p potential fixed regressors. The following probit mixed model is considered

$$P(Y_i = 1 \mid U, \beta) = p_i = \Phi(X_i^T \beta + Z_i^T U),$$

where Φ stands for the standard Gaussian cumulative distribution function, and X_i and Z_i for the fixed and random effect regressors associated to the ith observation. The parameter $\beta \in \mathbb{R}^p$ corresponds to the fixed-effect coefficients and the parameter U to the random-effect coefficients. X and Z are known design matrices associated with the fixed and random effects.

We consider K random effects, $U = (U_1^T, \cdots, U_K^T)^T$ where each U_l is a vector of size q_l, and $\sum_{l=1}^K q_l = q$. Following Albert and Chib [1993] and Lee et al. [2003], a vector of latent variables $L = (L_1, \ldots, L_n)^T$ is introduced, and we assume that the conditional distribution of L is Gaussian, written $L \mid U, \beta \sim \mathcal{N}_n(X\beta + ZU, I_n)$, with I_n the identity matrix. We then have

$$Y_i = \begin{cases} 1 & \text{if } L_i > 0 \\ 0 & \text{if } L_i < 0. \end{cases} \tag{1}$$

When the purpose is to select relevant variables among the p candidates, it is convenient to denote by γ the vector of latent variables indicating if a variable is selected or not; that is, $\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$. We then denote by β_γ the vector of all nonzero elements of β and by $\mathbf{X}_\gamma$ the design matrix with columns corresponding to the elements of γ that are equal to 1.

To complete the hierarchical model, a conventional prior distribution for $\beta_\gamma|\gamma$ is a d_γ-dimensional Gaussian distribution, with $d_\gamma = \sum_{j=1}^{p} \gamma_j$,

$$\beta_\gamma|\gamma \sim \mathcal{N}_{d_\gamma}(0, \Sigma_\gamma). \tag{2}$$

Concerning the prior covariance matrix Σ_γ, an attractive and standard choice is

$$\Sigma_\gamma = \tau(\mathbf{X}_\gamma' \mathbf{X}_\gamma)^{-1}. \tag{3}$$

Equations (2) and (3) together correspond to the g-prior distribution, first proposed by Zellner [1986], and commonly used since. The parameter $\tau > 0$ is referred to as the variable selection coefficient in Bottolo and Richardson [2010]. This g-prior replicates the covariance structure of the design and enables an automatic scaling based on the data. Moreover, it leads to simple expressions of the marginal likelihood. In the homoscedastic linear model with variance σ^2, it can be expressed as $\tau = g\sigma^2$. Up to the scalar τ, the prior covariance matrix is related to the Fisher Information Matrix (see for instance Chen and Ibrahim [2003]).

The choice of the variable selection coefficient τ can have a great influence on the variable selection process (see George and Foster [2000]) and has been considered by many authors. Some of them considered a fixed value for τ. For instance Smith and Kohn [1997] suggested to choose τ between 10 and 100. Another approach is the approach of George and Foster [2000], who developed empirical Bayes methods based on the estimation of τ from its marginal likelihood. Other authors proposed to put a hyper-prior distribution on τ, like Zellner and Siow [1980] that used an inverse-gamma distribution $\mathcal{IG}(1/2, n/2)$. But under the Zellner-Siow prior, marginal likelihoods are not available in closed forms, and approximations are necessary (see Bottolo and Richardson [2010]). Note also that the Zellner-Siow prior can be represented as a mixture of g-priors. Following this remark, Liang et al. [2008] proposed a new family of priors on τ, the *hyper-g prior family* which leads to new mixtures of g-priors while maintaining the computational tractability of the marginal likelihoods. Independently but in the same spirit, Cui and George [2008] suggested to put an inverse-gamma prior distribution on $(1 + \tau)$ (rather than on τ like Zellner and Siow), obtaining a family of priors on τ which contains the *hyper-g prior family* as a special case. Marin and Robert [2007] also proposed a way to use mixtures of g-prior for model selection.

In spite of the variety of all these works to choose the variable selection coefficient τ, a crucial problem remains with priors using the matrix $(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}$. Indeed, $\mathbf{X}_\gamma^T \mathbf{X}_\gamma$ should be invertible. However, there are two standard cases where $\mathbf{X}_\gamma^T \mathbf{X}_\gamma$ is singular:

- If the number of observations is lower than the number of variables in the model, $n < d_\gamma$.
- If some variables are linear combinations of others. In practice, even if $\mathbf{X}_\gamma^T \mathbf{X}_\gamma$ is theoretically invertible, some variables can be highly correlated and $\mathbf{X}_\gamma^T \mathbf{X}_\gamma$ can be computationally singular. It is often the case in genomic high-dimensional datasets for example. This

problem can also be encountered when several datasets are merged: some variables can be collinear or almost collinear if same variables were present into several datasets under different labels for instance.

In these cases the classical g-prior does not work. Concerning the first case, several authors proposed alternative priors. In case of linear models, Maruyama and George [2010] proposed a generalization of the g-prior. In case of probit models, Yang and Song [2010] proposed to replace the matrix $(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}$ in Σ_γ by its Moore Penrose's inverse (see alsoWest [2003]). However, the computation of the posterior distribution proposed by Yang and Song [2010] has a technical issue that do not permit the use of MCMC algorithm (see Baragatti and Pommeret [2011]). An other idea would be to avoid this first case by fixing the number of selected covariates at each iteration, as in Baragatti [2011]. It is a practical way when the purpose is to retain only few regressors, because it appeared computationally advantageous and it reduced the effect of the variable selection coefficient τ used in the g-prior. But the number of selected variables at each iteration must be arbitrarily fixed. Moreover, fixing the number of selected covariates is not a solution for the second case, as well as the priors proposed by Maruyama and George [2010] and Yang and Song [2010]. For linear model, Brown et al. [1998] proposed to work with transformations of X and Y, using a Singular Values Decomposition in the spirit of the ridge regression (see Marquardt [1970]). This numerical approach was extended for multivariate general linear models in Brown et al. [2002]. Always in the spirit of ridge regression, Gupta and Ibrahim [2007] proposed an extension of the g-prior, by introducing a ridge parameter. Their prior can be used in the two cases, but they did not really considered the second case in which some variables are linear combinations of others. In this paper we also propose a prior with a ridge parameter, which is slightly different from the prior of Gupta and Ibrahim [2007]. Besides, we suggest a way to choose the associated hyper-parameters: following the original idea of Zellner which is to keep the covariance structure of the design, we propose to keep the total variance of the data through the trace of $\mathbf{X}^T\mathbf{X}$. The model used is a probit mixed model, and the illustrations focus on the second case. Particularly, Affymetrix microarray experiment results from patients with breast cancer are studied.

The rest of the paper is organized as follows. In Section 2 we introduce the prior to be used in the probit mixed model and we suggest a choice for the hyper-parameters. The prior and full conditional distributions used in the algorithm are detailed. Section 3 outlines the algorithm. In Section 4 some experimental results on simulated and real datasets are given, and a sensitivity analysis is performed. Finally Section 5 discusses the method.

2 Introducing a ridge parameter

To complete the hierarchical model of Section 1, some prior assumptions have to be made on $U|D$, $\beta_\gamma|\gamma$, γ and D, with D a covariance matrix of dimension q.

2.1 Prior distribution of β with a ridge parameter

As previously explained, in case of singularity of the matrix $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$, the classical g-prior can not be used. We propose to introduce a ridge parameter, denoted $\lambda > 0$, by replacing in (3) the matrix $\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ by $\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I$. We get

$$\beta_\gamma|\gamma \sim \mathcal{N}_{d_\gamma}\big(0, (\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I)^{-1}\big) \qquad \text{with} \qquad d_\gamma = \sum_{j=1}^{p} \gamma_j. \tag{4}$$

We write

$$\Sigma_\gamma(\lambda) = (\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I)^{-1}. \tag{5}$$

With $\lambda > 0$, the matrix $\Sigma_\gamma(\lambda)$ is always of full rank. We obtain a modified form of the g-prior, which is a compromise between independence and instability. Indeed, for large values of λ and τ, $\Sigma_\gamma(\lambda)$ is close to a diagonal matrix that coincides with the conditional independent case. On the opposite, for small values of λ and τ, the term $\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ prevails and the inverse of $\tau^{-1}\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I$ will be instable if $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ is singular. In that case, the prior distribution (4) is close to the g-prior case. In comparison, the prior of Gupta and Ibrahim [2007] uses $\Sigma_\gamma(\lambda) = \tau(\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I)^{-1}$ (with an additional parameter σ^2 because they focused on the linear case). Note that the classical g-prior corresponds to a special case of (5), with $\lambda = 0$.

2.2 Calibrating hyper-parameters

Following Zellner [1986], our purpose is to use the design to calibrate the covariance of β_γ with a ridge parameter. Write $\Sigma_\gamma(0) = \tau_0(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}$, with τ_0 the fixed hyper-parameter used in this classical prior. Using $\Sigma_\gamma(\lambda)$ instead of $\Sigma_\gamma(0)$ corresponds to introducing a perturbation in the classical g-prior. As this classical prior gives good results when the matrix $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ is invertible, we choose λ and τ such that $\Sigma_\gamma(\lambda)^{-1}$ and $\Sigma_\gamma(0)^{-1}$ are as close as possible. Since $tr(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)$ represents the total variance (up to a normalization) explained by the selected covariates, the constraint used is

$$tr\Big(\Sigma_\gamma(0)^{-1}\Big) = tr\Big(\Sigma_\gamma(\lambda)^{-1}\Big),$$

which yields

$$\tau = \tau_0\Big[1 + \frac{\lambda p \tau_0}{tr(\mathbf{X}_\gamma^T\mathbf{X}_\gamma) - \lambda p \tau_0}\Big].$$

Concerning the choice of λ, in order to take into account the number p of covariates and to reduce the effect of the ridge factor, we suggest taking $\lambda = 1/p$, getting

$$\tau = \tau_0\Big[1 + \frac{\tau_0}{tr(\mathbf{X}_\gamma^T\mathbf{X}_\gamma) - \tau_0}\Big].$$

But the vector γ can be different between two iterations of the algorithm. Therefore we propose to use the complete design matrix $\mathbf{X}$ instead of $\mathbf{X}_\gamma$, yielding

$$\tau = \tau_0\Big[1 + \frac{\tau_0}{tr(\mathbf{X}^T\mathbf{X}) - \tau_0}\Big]. \tag{6}$$

In practice, the user has to choose only the parameter τ_0, λ and τ are then obtained by $1/p$ and (6). Following Smith and Kohn [1997], τ_0 can be chosen between 10 and 100. In section 4.3, the sensitivity analysis will assess the influence of the values of these hyper-parameters λ and τ.

2.3 Other prior distributions

The γ_j are assumed to be independent Bernoulli variables, with

$$P(\gamma_j = 1) = \pi_j, \qquad 0 \leq \pi_j \leq 1 \qquad j = 1, \ldots, p \tag{7}$$

If we do not want to use prior knowledge to favor any variables, we put $\pi_j = \pi, \forall j \in \{1, \ldots, p\}$. The vector of coefficients associated with the random effects is assumed to be Gaussian and centered, with covariance matrix D:

$$U|D \sim \mathcal{N}_q(0, D). \tag{8}$$

This definition allows three cases to be distinguished:

General case: No structure is assumed for the variance-covariance matrix D, its prior distribution is an Inverse-Wishart $\mathcal{W}^{-1}(\Psi, m)$.

Case of a block-diagonal matrix D: The different random effects are assumed independent. The vectors of coefficients associated with each random effect have Gaussian prior distributions:

$$U_l \mid A_l \sim \mathcal{N}_{q_l}(0, A_l), \quad l = 1, \ldots, K,$$

where the A_l are symmetric design matrices of dimension q_l. D is a block-diagonal matrix denoted by $diag(A_1, \ldots, A_K)$. The prior distributions for each A_l are Inverse-Wishart $\mathcal{W}^{-1}(\Psi, m)$.

Case of a diagonal matrix D: $D = diag(A_1, \ldots, A_K)$ where $A_l = \sigma_l^2 I_{q_l}$, $l = 1, \ldots, K$ and I_{q_l} the identity matrix. The prior distributions for the σ_l^2 are then Inverse Gamma $\mathcal{IG}(a, b)$ (b denoting the scale parameter).

2.4 Conditional distributions

The posterior distribution of γ is of particular interest for the variable selection problem. An idea is to use a Gibbs sampler to explore this posterior distribution and to search for high probability γ values. Therefore, we must be able to simulate from all of the full conditional distributions (simplified by the hierarchical structure): $f(L \mid Y, \beta, U)$, $f(\beta \mid L, U, \gamma)$, $f(U \mid L, \beta, D)$, $f(\gamma \mid L, U, \beta)$ and $f(D \mid U)$. The advantage of the ridge approach is that these posterior distributions are available in closed forms.

- Full conditional distribution of L.

$$\begin{aligned} L_i|\beta, U, Y_i = 1 &\sim \mathcal{N}(X_i^T\beta + Z_i^T U, 1) \text{ left truncated at } 0 \\ L_i|\beta, U, Y_i = 0 &\sim \mathcal{N}(X_i^T\beta + Z_i^T U, 1) \text{ right truncated at } 0 \end{aligned} \tag{9}$$

- Full conditional distribution of β_γ. Given γ, we know which elements of β are not null. So we focus on the generation of the non null elements of β_γ.

$$\beta_\gamma|L, U, \gamma \sim \mathcal{N}_{d_\gamma}(V_\gamma \mathbf{X}_\gamma^T(L - ZU), V_\gamma), \tag{10}$$

where

$$V_\gamma = \Big[\frac{(1+\tau)}{\tau}\mathbf{X}_\gamma^T\mathbf{X}_\gamma + \lambda I\Big]^{-1}.$$

- Full conditional distribution of U. Defining $W = (Z^T Z + D^{-1})^{-1}$, we have

$$U|L,\beta,D \sim \mathcal{N}_q(WZ^T(L-\mathbf{X}\beta),W), \tag{11}$$

- Full conditional distribution of D.
  General case: The full conditional distribution of D is an Inverse-Wishart:

$$D \mid U \quad \sim \quad \mathcal{W}^{-1}(UU^T + \Psi, m+1). \tag{12}$$

  Case of a block-diagonal matrix D: $D = diag(A_1,\dots,A_K)$. The full conditional distribution of A_l ($\forall l = 1,\dots,K$) is an Inverse-Wishart:

$$A_l \mid U_l \quad \sim \quad \mathcal{W}^{-1}(U_l U_l^T + \Psi, m+1). \tag{13}$$

  Case of a diagonal matrix D: $D = diag(A_1,\dots,A_K)$, and $\forall l = 1,\dots,K$, $A_l = \sigma_l^2 I_{q_l}$. The full conditional distribution of σ_l^2 is an Inverse-Gamma:

$$\sigma_l^2 \mid U_l \quad \sim \quad \mathcal{IG}\Big(\frac{q_l}{2}+a, \big(\frac{1}{2}U_l^T U_l + b\big)\Big). \tag{14}$$

- Full conditional distribution of γ.

$$\begin{aligned} f(\gamma|\beta_\gamma, L, U) \quad &\propto \quad |\Sigma_\gamma(\lambda)|^{-1/2}\prod_{j=1}^{p}\pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j} \qquad (15)\\ &\times \quad (2\pi)^{-\frac{d_\gamma}{2}}\exp\Big[-\frac{1}{2}\big((ZU-L)^T\mathbf{X}_\gamma\beta_\gamma + \beta_\gamma^T\mathbf{X}_\gamma^T(ZU-L) + \beta_\gamma^T V_\gamma^{-1}\beta_\gamma\big)\Big], \end{aligned}$$

  with $d_\gamma = \sum_{j=1}^p \gamma_j$.

We want to simulate γ from the distribution (15) using a Metropolis-Hastings (MH) algorithm and not component by component, because it is computationally advantageous for a very large number of variables, see Baragatti [2011]. However, the full conditional distribution of γ cannot be directly simulated using a MH algorithm, since it depends on the actual value of β_γ. Following Lee et al. [2003] and Baragatti [2011], we use the grouping (or blocking) technique of Liu [1994] to eliminate the nuisance parameter β_γ. The idea is to group the parameters β_γ and γ, so we will be interested in the full conditional distribution of $(\beta_\gamma,\gamma) \mid L, U$. This technique improves the algorithm and facilitates the convergence of the Markov chain, see Liu [1994] and van Dyk and Park [2008]. As we have

$$f(\beta_\gamma,\gamma \mid L,U) \propto f(\gamma \mid L,U)f(\beta_\gamma \mid \gamma, L, U),$$

we note that simulating from the full conditional distribution $(\beta_\gamma,\gamma) \mid L,U$ is equivalent to simulating γ from its full conditional distribution integrated on β_γ, then simulating β_γ from its

full conditional distribution. The "integrated distribution" for γ will not depend anymore on the nuisance parameter β_γ.
Having integrated β_γ out in equation 15, we obtain

$$\begin{aligned} f(\gamma|L,U) &\propto \frac{|V_\gamma|^{1/2}}{|\Sigma_\gamma(\lambda)|^{1/2}} \exp\Big[-\frac{1}{2}((L-ZU)^T(I-\mathbf{X}_\gamma V_\gamma \mathbf{X}_\gamma^T)(L-ZU)\Big] \\ &\times \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}. \end{aligned} \tag{16}$$

Remark: The influence of τ appears here through the ratio $R^{1/2} = \left(\frac{|V_\gamma|}{|\Sigma_\gamma|}\right)^{1/2}$. We can see that

$$\begin{cases} \text{if } \tau \to \infty, & R \to |\mathbf{X}_\gamma^T \mathbf{X}_\gamma + \lambda I|^{-1}, \\ \text{if } \tau \to 0, & R \to 1. \end{cases}$$

3 Metropolis-within-Gibbs algorithm

3.1 A Metropolis-Hastings step to simulate γ

At iteration $(i+1)$ of the Metropolis-Hastings algorithm, a candidate γ^* is proposed from $\gamma^{(i)}$. Using a symmetric transition kernel, the acceptance rate is

$$\rho(\gamma^{(i)},\gamma^*) = \min\left\{1, \frac{f(\gamma^*|L,U)}{f(\gamma^{(i)}|L,U)}\right\},$$

with

$$\begin{aligned} \frac{f(\gamma^*|L,U)}{f(\gamma^{(i)}|L,U)} &= \left(\frac{|V_{\gamma^*}\Sigma_{\gamma^{(i)}}|}{|\Sigma_{\gamma^*}V_{\gamma^{(i)}}|}\right)^{1/2} \exp\Big\{-\frac{1}{2}(L-ZU)^T(\mathbf{X}_{\gamma^i}V_{\gamma^{(i)}}\mathbf{X}_{\gamma^{(i)}}^T - \mathbf{X}_{\gamma^*}V_{\gamma^*}\mathbf{X}_{\gamma^*}^T)(L-ZU)\Big\} \\ &\times \prod_{j=1}^{p}\left(\frac{\pi_j}{1-\pi_j}\right)^{\gamma_j^*-\gamma_j^{(i)}}, \qquad \text{if} \qquad \forall j \in \{1,\ldots,p\} \quad \pi_j = \pi. \end{aligned} \tag{17}$$

The simplest way to have a symmetric transition kernel is to propose a γ^* which corresponds to $\gamma^{(i)}$ in which r components have been randomly changed (see Chipman et al. [2001] and George and McCulloch [1997]).

Remark: The influence of τ appears via the ratio $Q^{1/2} = \left(\frac{|V_{\gamma^*}\Sigma_{\gamma^{(i)}}|}{|\Sigma_{\gamma^*}V_{\gamma^{(i)}}|}\right)^{1/2}$ that satisfies:

$$\begin{cases} \text{if } \tau \to \infty, & Q \to |\mathbf{X}_{\gamma^*}^T\mathbf{X}_{\gamma^*} + \lambda I| \times |\mathbf{X}_{\gamma^i}^T\mathbf{X}_{\gamma^i} + \lambda I|^{-1}, \\ \text{if } \tau \to 0, & Q \to 1. \end{cases}$$

3.2 Complete algorithm

The Metropolis-within-Gibbs sampler (Roberts and Rosenthal [2006]) modified by the grouping technique of Liu generates a sequence:

$$\gamma^{(1)}, \beta_\gamma^{(1)}, D^{(1)}, L^{(1)}, U^{(1)}, \ldots\ldots, \gamma^{(b+m)}, \beta_\gamma^{(b+m)}, D^{(b+m)}, L^{(b+m)}, U^{(b+m)}.$$

The sequence of the $\gamma^{(t)}$, which is of interest for the variable selection problem, is embedded in this "Gibbs sequence".

Algorithm:
Starting with initial values $\gamma^{(0)}, \beta^{(0)}, D^{(0)}, L^{(0)}, U^{(0)}$. At iteration $t+1$:

1. Simulate $\gamma^{(t+1)}$ from $f(\gamma \mid L^{(t)}, U^{(t)})$ (see 16), using the Metropolis-Hasting step. Given $\gamma^{(t)}, L^{(t)}, U^{(t)}$, k iterations of the Metropolis-Hastings algorithm are performed (k arbitrarily fixed). The Metropolis-Hastings step begins with $\gamma^{(t)}$ as an initial value. Then at each iteration $i+1$:
   (a) Generate the γ^* candidate, by changing r components of $\gamma^{(i)}$.
   (b) Take
   $$\gamma^{(i+1)} = \begin{cases} \gamma^* & \text{with probability} \quad \rho(\gamma^{(i)}, \gamma^*), \qquad \text{see (17)} \\ \gamma^{(i)} & \text{with probability} \quad 1-\rho(\gamma^{(i)}, \gamma^*) \end{cases}$$
   $\gamma^{(t+1)}$ will be the $\gamma^{(k)}$ obtained at the k^{th} iteration of the Metropolis-Hastings algorithm.
2. Simulate $\beta_\gamma^{(t+1)}$ from $f(\beta_\gamma \mid L^{(t)}, U^{(t)}, \gamma^{(t+1)})$ (see (10)).
3. Simulate $D^{(t+1)}$ from $f(D \mid U^{(t)})$ (see (12), (13) or (14)).
4. Simulate $L^{(t+1)}$ from $f(L \mid Y, \beta^{(t+1)}, U^{(t)})$ (see (9)).
5. Simulate $U^{(t+1)}$ from $f(U \mid L^{(t+1)}, \beta^{(t+1)}, D^{(t+1)})$ (see (11)).

The number of iterations is $b+m$, where b corresponds to the burn-in period and m to the observations from the posterior distributions. For selection of variables, the sequence $\{\gamma^{(t)} = (\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}), t = b+1, \ldots, b+m\}$ is used. The most relevant variables for the regression model are those which are supported by the data and prior information. Thus they are those corresponding to the γ components with higher posterior probabilities, and can be identified as the γ components that are most often equal to 1. To decide which variables should be finally selected after a run, we suggest to use a box-plot of the number of iterations during which variables were selected. Usually, for each run a reasonable number of variables distinguishable from others can be selected by fixing a threshold: if a variable has been selected during a number of iterations which is higher than this threshold, then the variable is kept in the final selection.

4 Experimental results

4.1 Simulated data

We simulated 200 binary observations and 300 variables, the observations being obtained using a probit mixed model with 5 of these variables and one random effect of length 4. Among the

300 variables, 280 were generated from a uniform on $[-5, 5]$ and denoted by $V1, \dots, V280$. Then 10 variables denoted by $V281, \dots, V290$ were build to be collinear to the first 10 variables, with a factor 2: for instance $V282 = 2 \times V2$. One variable was build to be a linear combination of $V1$ and $V2$ ($V291 = V1 + V2$), and another was build to be a linear combination of $V3$ and $V4$ ($V292 = V3 - V4$). Finally, 8 variables were build to be linear combinations of variables 5 to 12 and variables 13 to 20 (for instance $V293 = V5 + V13$). The five variables used to generate the binary observations were the first five: $V1, V2, V3, V4$ and $V5$. The vector of coefficients associated with these variables was $\beta = (1, -1, 2, -2, 3)$. The first 100 observations were part of the training set, and the last 100 were part of the validation set. In the training and the validation sets, 25 observations were associated with each component of the random effect, whom vector of coefficients was $U = (-3, -2, 2, 3)$. We had only one random effect and the different components were supposed independent, hence we put $D = \sigma^2 I_3$ with an inverse-gamma prior $\mathcal{IG}(a, b)$ for σ^2.
The objective was to assess the behavior of the proposed method when some variables are linear combinations of others, and to compare it to the case where no variable is linear combination of others. Therefore we performed 10 runs of the algorithm using only the first 280 variables, and 10 runs using the 300 variables. In these two cases and for each run the same parameters were used: 5 variables were initially selected, one component of γ was proposed to be changed at each iteration of the Metropolis-Hastings step, the prior of σ^2 was a $\mathcal{IG}(1, 1)$, $\pi_j = 5/280$ for all j when 280 variables were kept, $\pi_j = 5/300$ for all j when 300 variables were kept, 4000 iterations were performed after a burn-in period of 1000 iterations, and each Metropolis-Hastings step consisted of 500 iterations. We decided to choose $\tau_0 = 50$, which is a standard choice, see Smith and Kohn [1997] for instance. The parameters λ and τ were then chosen as explained in 2.2 and using (6), yielding $\lambda = 1/280$ and $c = 50.01075$ when using 280 variables, and $\lambda = 1/300$ and $c = 50.00885$ when using 300 variables.

A final selection was performed for each of the 20 runs, by taking the variables which were selected the most often during the run. Boxplots were used to determine the threshold above which the variables are kept. Figure 1 presents two boxplots: one corresponding to a run with only the first 280 variables, and one corresponding to a run with the 300 variables.

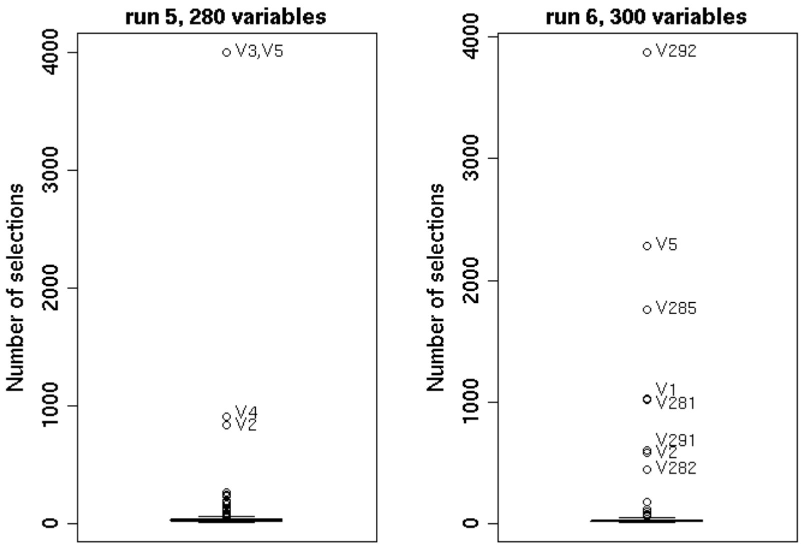

Figure 1: Boxplots of the number of selections of a variable after the burn-in period. A point represents a variable (or several variables if they have been selected the same number of times). The left boxplot corresponds to the run 5 with 280 variables: there is a gap between the variables $V2, V3, V4$ and $V5$ selected in more than 500 iterations and the others, hence we selected these four variables. The right boxplot corresponds to the run 6 with 300 variables: there is a gap between the variables selected in more than 400 iterations and the others, hence we selected these eight variables

Table 1 gives the variables kept in the final selections of the 10 runs with the first 280 variables, and of the 10 runs with 300 variables.

Variables	Number of selections among the 10 runs with 280 variables	Number of selections among the 10 runs with 300 variables
$V1$	0	10
$V2$	9	8
$V3$	10	2
$V4$	5	0
$V5$	10	10
$V281 = 2 \times V1$	Not available	10
$V282 = 2 \times V2$		9
$V283 = 2 \times V3$		3
$V284 = 2 \times V4$		0
$V285 = 2 \times V5$		10
$V291 = V1 + V2$		7
$V292 = V3 - V4$		10

Table 1: Number of final selections among the 10 runs with the first 280 variables and among the 10 runs with 300 variables, for the variables $V1, V2, V3, V4, V5$ and linear combinations of these variables. No other variable was present on the final selections.

Among the runs with the first 280 variables, 3 among the 5 variables used to generate the data were in the final selection of almost all runs, and the variables $V4$ was in the final selection of half of the runs. Notice that $V1$ was in none of the final selections. Among the runs with 300 variables, the variables $V1, V2, V3$ and $V5$ were present in most of the final selections, directly or indirectly through linear combinations. Contrarily to the runs with 280 variables, the variables $V4$ or $V284$ were in none of the final selections, while the variables $V1$ and $V281$ were in all the final selections. Concerning $V4$, it was indirectly in all the final selections through $V292$, which is a linear combination of $V3$ and $V4$. Eventually, the final selections of the runs with 300 variables appeared as relevant as the final selections of the runs with 280 variables, despite the fact that some variables were linear combinations of others. Note that we obtained similar results with only 500 burn-in iterations and 500 post burn-in iterations, except that the variable $V4$ was in none of the final selections.

To assess the relevance of the final selections, predictions were performed. Sensitivity and specificity are presented in Table 2.

Variables selected among 280			Variables selected among 300		
Variables	Sensitivity	Specificity	Variables	Sensitivity	Specificity
$V2, V3, V5$	0.87	0.89	$V281$, $V282$ $V283$, $V285$ and $V292$	0.94	0.89
$V2, V3, V4, V5$	0.93	0.96			
True model: $V1, V2, V3, V4, V5$				0.94	0.89

Table 2: Sensitivity and specificity on the validation dataset.

Concerning the runs with the first 280 variables, the variables $V2, V3$ and $V5$ were in almost all final selections. Fitting a probit mixed model on the training set with variables $V2, V3$ and $V5$, and making predictions on the validation set, we obtained 12 misclassifications among 100. This result was good, despite the fact that $V1$ and $V4$ were not taken into account. $V4$ was in half of the final selections, hence we fitted a probit mixed model on the training set with variables $V2, V3, V4$ and $V5$, and made predictions on the validation set: we obtained 6 misclassifications among 100. For comparison, using the five variables used to generate the data, we obtained 8 misclassifications. Therefore the results of the algorithm were good enough. We then did predictions using the variables in final selections of the runs with 300 variables. Notice that we can not fit a model with all the variables in final selections, because some of them are linear combinations of others. For instance, we can not use $V1$ and $V281$ together. We fit a probit mixed model on the training set with variables $V281, V282, V283, V285$ and $V292$, and made predictions on the validation set: we obtained 8 misclassifications among 100. It was as good as with the five variables used to generate the data.

The number of components of γ equal to 1 (corresponding to d_γ) can vary from one iteration to another. Indeed, during the Metropolis-Hastings step, at each iteration a new γ vector is proposed. Figure 2 shows, for the 10 runs with 300 variables, the number of iterations of the runs associated with a number of selected variables from 1 to 15.

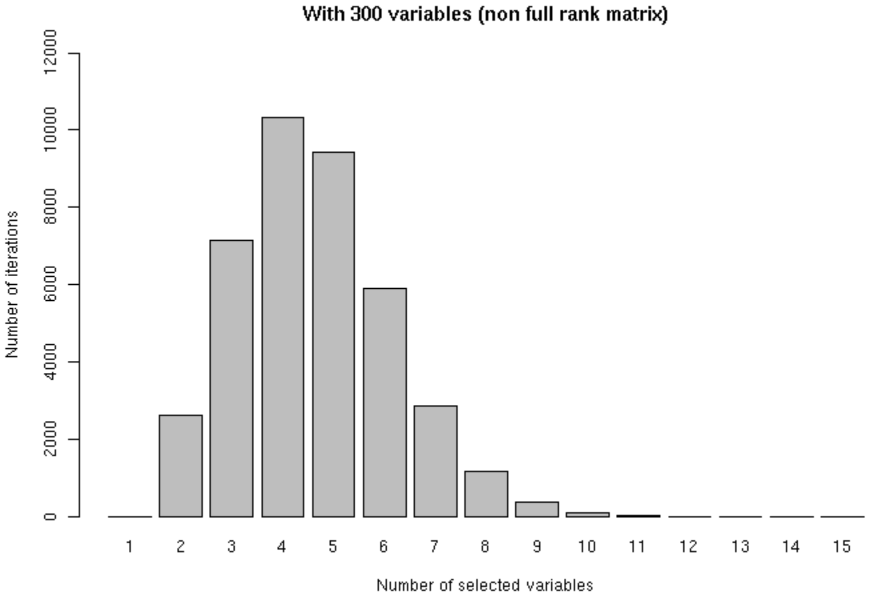

Figure 2: Number of iterations of the runs associated with a number of selected variables from 1 to 15. For the 10 runs, there were a total of 40000 post burn-in iterations.

Similar results were obtained for the 10 runs with the first 280 variables. The number of variables selected at each iteration never became larger than 13, hence the case $p > n$ have not be encountered. Table 3 gives the number of variables in the final selections of the 10 runs with 280 variables, and of the 10 runs with 300 variables.

	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
With 280 variables	4	4	3	3	4	3	2	3	4	4
With 300 variables	5	6	8	8	8	8	10	9	8	9

Table 3: Number of variables in the final selections of the 10 runs with 280 variables, and of the 10 runs with 300 variables.

The number of variables in the final selections of the 10 runs with 280 variables appeared to be generally lower than this number in the final selections of the 10 runs with 300 variables.

4.2 Illustrations through real data

As an illustration, Affymetrix microarray experiment results from patients with breast cancer were used. For that, we consider data used in Baragatti [2011], see there for more details. Briefly, the patients come from three different hospitals, and the objective was to select some variables (probesets) which are indicative of the activity of the estrogen receptor (ER) gene in

breast cancer. The hospital was considered as a random effect in the model, thus accounting for the different experimental conditions between the three hospitals. For each patient, the expressions of 275 probesets were kept, among which some were known to be relevant to explain the ER status (corresponding to variables 148, 260, 263 and 273). We used a training set made of 100 patients, and a validation set of 88 patients. In order to have a potentially singular $\mathbf{X}_\gamma^T \mathbf{X}_\gamma$ matrix, we added three variables to the data matrix $\mathbf{X}$. These variables were linear combinations of the known relevant variables, hence $\mathbf{X}$ was no more of full rank: $V276 = 2 \times V148$, $V277 = -V260$ and $V278 = V263 + V273$. We had only one random effect, which corresponded to the different hospitals. The hospitals are supposed independent, hence we put $D = \sigma^2 I_3$ with an inverse-gamma prior $\mathcal{IG}(a, b)$ for σ^2.
We performed 10 runs of the algorithm using only the first 275 variables, and 10 runs using all the 278 variables. In these two cases and for each run the same 100 patients and the same parameters were used: 5 variables were initially selected, one component of γ was proposed to be changed at each iteration of the Metropolis-Hastings step, the prior of σ^2 was a $\mathcal{IG}(1, 1)$, $\pi_j = 5/275$ for all j when 275 variables were kept, $\pi_j = 5/278$ for all j when 278 variables were kept, 4000 iterations were performed after a burn-in period of 1000 iterations, and each Metropolis-Hastings step consisted of 500 iterations. As in the previous illustration we chose $\tau_0 = 50$. The parameters λ and τ were then chosen as explained in 2.2 and using (6), yielding $\lambda = 1/275$ and $c = 50.0009$ when using 275 variables, and $\lambda = 1/278$ and $c = 50.00088$ when using 278 variables.

A final selection was performed for each of the 20 runs, by taking the variables which were selected the most often during the run. Boxplots were used to determine the threshold above which the variables are kept. Figure 3 presents a boxplot of a run with 275 variables and a boxplot of a run with 278 variables.

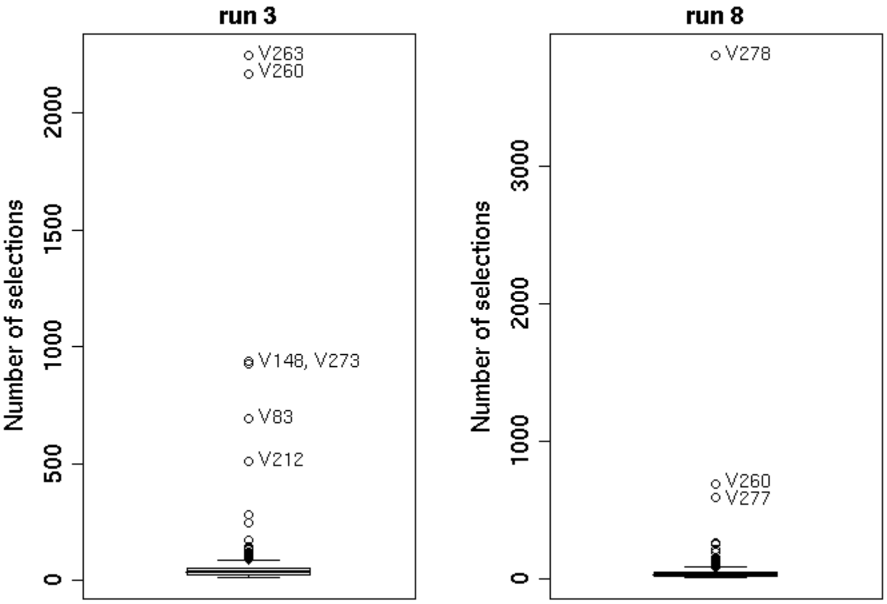

Figure 3: Boxplots of the number of selections of a variable after the burn-in period. The left boxplot corresponds to the run 3 with 275 variables: there is a gap between the variables selected in more than 500 iterations and the others, hence we selected these six variable. The right boxplot corresponds to the run 8 with 278 variables: there is a gap between the variables $V278, V260$ and $V277$ selected in more than 500 iterations and the others, hence we selected these three variables.

Table 4 gives the variables kept in the final selections of the 10 runs with the first 275 variables, and of the 10 runs with 278 variables.

Variables	Corresponding probesets	Number of selections among the 10 runs with 275 variables	Number of selections among the 10 runs with 278 variables
$V260$	228241_at	10	3
$V273$	205862_at	9	0
$V148$	209604_s_at	5	1
$V263$	228554_at	10	0
$V83$	203628_at	7	0
$V66$	202088_at	1	0
$V212$	215157_x_at	1	0
$V277 = -V260$	collinearity	Not available	3
$V278 = V263 + V273$	linear combination		10

Table 4: Number of final selections among the 10 runs with the first 275 variables and among the 10 runs with 278 variables, for the different variables and linear combinations. No other variable was present on the final selections.

Concerning the runs with the first 275 variables, three of the most relevant probesets were in the final selections of most of them: $V260$, $V273$ and $V263$. Moreover, $V148$ was in half of the final selections. Concerning the runs with 278 variables, the variable $V278$ which is a linear combination of $V263$ and $V273$ was in all of the final selections. We noticed that the variables $V260$ and $V277$ were in the same final selections, and that $V83$ was selected during runs with a non singular X matrix and not during runs with a singular X matrix. As in the previous example, the final selections of the runs with 278 variables appeared as relevant as the final selections of the runs with 275 variables, despite the fact that some variables were linear combinations of others. We obtained similar results with only 500 burn-in iterations and 500 post burn-in iterations, except that two of the most relevant variables were in most of the final selections, and not three ($V260$ and $V263$).

Predictions were also performed. Table 5 contains sensitivity and specificity. Concerning the runs with 275 variables, we fit a probit mixed model on the training set with variables $V260, V273$ and $V263$, and made predictions on the validation set: we obtained 4 misclassifications among 88. For comparison, using the four relevant variables $V260, V273, V148$ and $V263$, we obtained 3 misclassifications. Therefore the results of the algorithm were good enough. We then did predictions using the variables in final selections of the runs with 278 variables. Fitting a probit mixed model on the training set with only the variable $V278$ and making predictions on the validation set, we obtained 8 misclassifications among 88. Fitting a probit mixed model on the training set with the variables $V278$ and $V277$ and making predictions on the validation set, we obtained 3 misclassifications among 88, hence the results were quite good.

Variables selected among 275			Variables selected among 278		
Variables	Sensitivity	Specificity	Variables	Sensitivity	Specificity
$V260, V273, V263$	0.92	1	$V278$	0.87	0.97
$V260, V273, V148, V263$	0.94	1	$V278$, $V277$	0.94	1

Table 5: Sensitivity and specificity on the validation dataset.

Concerning the number of components of γ equal to 1 (corresponding to d_γ) during the iterations of the runs, figures quite similar to Figure 2 were obtained for the 10 runs with 275 variables, and for the 10 runs with 278 variables. The mode of the barplot was still for 4 variables selected in an iteration. Furthermore, as in the previous example the number of variables selected at each iteration never became larger than 14, therefore we have not been in the case $p > n$. Table 6 gives the number of variables in the final selections of the 10 runs with 275 variables, and of the 10 runs with 278 variables.

	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
With 275 variables	4	4	6	5	3	5	3	3	5	5
With 278 variables	1	1	1	2	1	3	1	3	3	1

Table 6: Number of variables in the final selections of the 10 runs with 275 variables, and of the 10 runs with 378 variables.

In opposition to the previous example, the number of variables in the final selections of the 10 runs in case of no linear combination appeared to be generally higher than this number in the final selections of the 10 runs in case of linear combinations.

4.3 Sensitivity analysis

Concerning the variable selection coefficient τ, the method of variable selection without the ridge parameter is not sensitive to its value (see Baragatti [2011]), but it is mainly due to the fact that the number of variables selected at each iteration of this algorithm was fixed. It is no more the case for the algorithm proposed in this paper, therefore it seems necessary to assess its sensitivity to this parameter. Similarly, it seems necessary to assess its sensitivity to the parameter λ. Indeed, we suggested a way to choose τ and λ, but it is interesting to study the behavior of the algorithm if these parameters are chosen more arbitrarily. As a consequence, we looked at the influence of τ_0 when τ is chosen from τ_0 and λ is chosen as proposed in Section 2.2, as well as the influences of τ and λ when they are chosen more arbitrarily.
We also studied the behavior of the algorithm when the value of the π_j, the prior distribution parameters of σ^2 and the number of iterations vary.
For this sensitivity study we used the example with simulated data (Section 4.2) with 300 variables, and the different values for the parameters used are presented in Table 7. In this table, the number of relevant variables in the final selections of the runs are given, the relevant variables being $V1, V2, V3, V4, V5, V281$, $V282, V283, V284, V285, V291$ and $V292$. The sensitivity was assessed by using the relative weighted consistency measure of Somol and Novovicova [2008], denoted by CW_{rel}. It is a measure evaluating how much subsets of selected variables for several runs overlap, and it shows the relative amount of randomness inherent in the concrete variable selection process. It takes values between 0 and 1, where 0 represents the outcome of completely random occurrence of variables in the selected subsets and 1 indicates the most stable variable selection outcome possible.

Run	c_0	c	λ	Value of π_j $\forall j$	Prior for σ^2	Iterations post burn-in (burn-in)	Nb of relevant variables	$\mathcal{S}$
1	10	10.00035 (6)	$1/p = 1/300$	5/300	$\mathcal{IG}(1,1)$	4000 (1000)	3	0.857
2	50	50.00885 (6)					8	
3	100	100.0354 (6)					8	
4	1000	1003.553 (6)					8	
5	10000	10367.03 (6)					8	
6	(6) non used	100	$1/p = 1/300$	5/300	$\mathcal{IG}(1,1)$	4000 (1000)	8	0.8
7			$100/p = 1/3$				8	
8			1				8	
9			10				8	
10			100				3	
11	(6) non used	10	$1/p = 1/300$	5/300	$\mathcal{IG}(1,1)$	4000 (1000)	5	0.348 (0.639 without run 13)
12		10	10				3	
13		1000	$1/p = 1/300$				0	
14		1000	10				8	
15		100	$100/p = 1/3$				8	
16	100	100.0354 (6)	1/p=1/300	5/300	$\mathcal{IG}(1,1)$	4000 (1000)	8	0.848
17				50/300			12	
18				100/300			12	
19	100	100.0354 (6)	1/p=1/300	5/300	$\mathcal{IG}(1,1)$	4000 (1000)	8	1
20					$\mathcal{IG}(2,5)$		8	
21					$\mathcal{IG}(5,2)$		8	
22	100	100.0354 (6)	1/p=1/300	5/300	$\mathcal{IG}(1,1)$	500 (500)	8	1
23						4000 (1000)	8	
24						40000 (10000)	8	

Table 7: Parameters of the runs for the sensitivity study and associated relative weighted consistency measure of Somol and Novovicova CW_{rel}. For each run, 5 variables are initially selected, one component of γ is proposed to be changed at each iteration of the Metropolis-Hastings step and each Metropolis-Hastings step consists of 500 iterations.

The algorithm was generally not sensitive to the values of the hyper-parameters, since most of the relevant variables were usually finally selected. When three variables only were finally selected, they were variables enabling us to fit models with good predictions. The boxplots obtained were often similar to the right boxplot of Figure 1. In particular, the algorithm was not sensitive to the values of τ and λ. However, the run 13 is noticeable, as no variable could be really distinguished from others, and none of the top-ranked variables was a relevant one, see Figure 4. This run corresponds to a large τ and a small λ. Note that the run 14 with large τ and large λ gave good results, even if the prior covariance of β_γ was then close to the identity. The runs 17 and 18 are also noticeable, as all relevant variables were finally selected, see Figure 4. They correspond to high values of π_j, and the cost for these relevant runs was longer computational times. Finally, we observed that the values of τ and π_j play role in the number of variables selected at each iteration of the algorithm. The value of τ modified the distribution of this number, see Figure 5. Besides, this number increased with the value of π_j, see Figure 6. However, even if the number of variables selected at each iteration of the algorithm was high, it did not influence the final selections of the runs, and it did not influence the number of variables which were distinguishable from others.

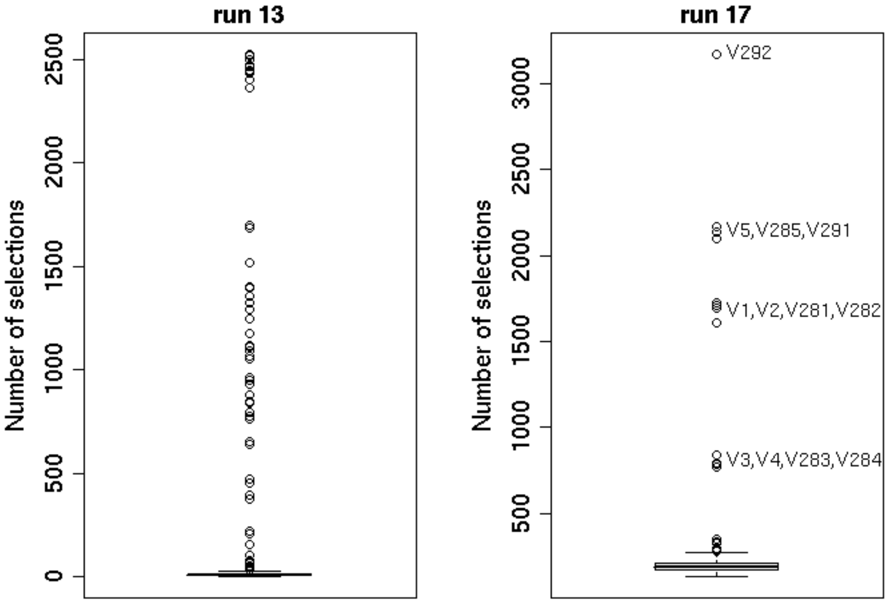

Figure 4: Boxplot of the number of selections of a variable after the burn-in period, for two runs with 300 variables. The left boxplot corresponds to the run 13: no variable distinguishes itself from others, and none of the top-ranked variables is a relevant one. The right boxplot corresponds to the run 17: the 12 relevant varaibles have been selected in more than 500 iterations.

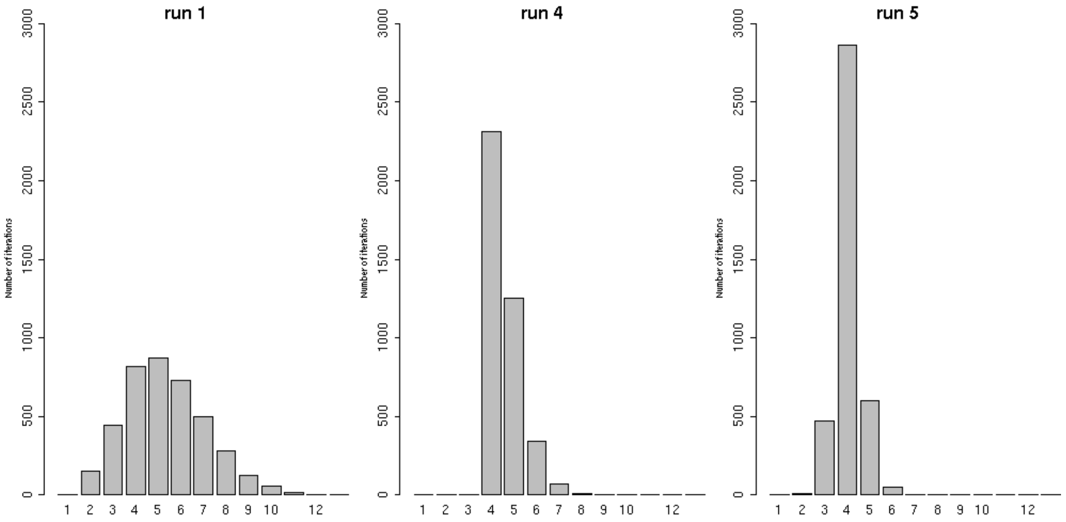

Figure 5: Number of iterations of the runs 1,4 and 5 associated with a number of selected variables from 1 to 14. For each run, there were 4000 post burn-in iterations.

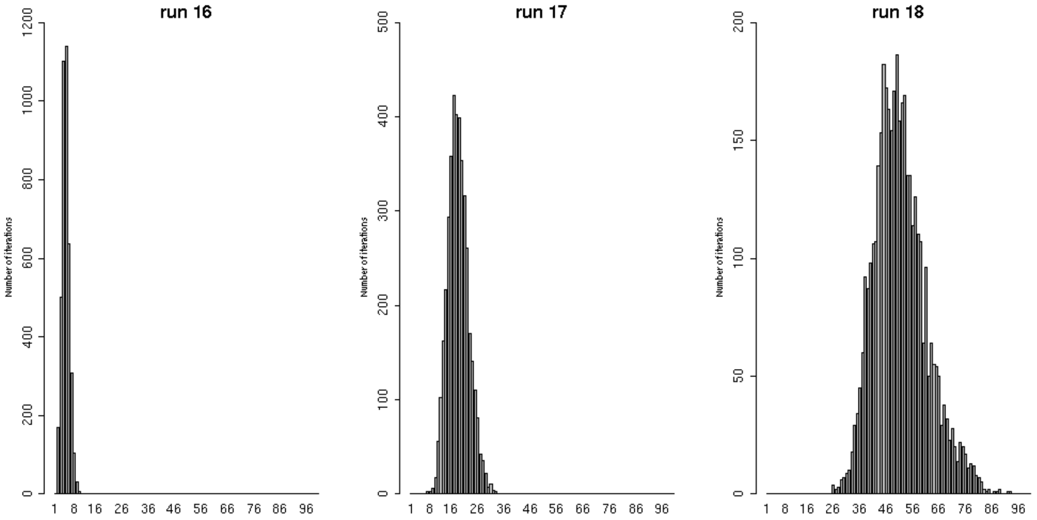

Figure 6: Number of iterations of the runs 16,17 and 18 associated with a number of selected variables from 1 to 100. For each run, there were 4000 post burn-in iterations.

5 Discussion

Classical bayesian variable selection methods often propose the use of the g-prior of Zellner. This prior can not be used if $p > n$, or if some variables are linear combinations of others. In particular, this last case can occur when several datasets with common covariates are merged. The prior for β_γ proposed in this manuscript is a possible alternative, as well as the prior proposed by Gupta and Ibrahim [2007]. In our prior the parameters τ and λ can be chosen independently, and in this case the parameter τ does not influence the coefficient of the identity matrix. On the opposite, in the prior of Gupta and Ibrahim, the parameter τ necessarily has an influence on this coefficient. Using our prior, a way to jointly choose τ and λ was suggested and the results obtained on simulated data and on a real dataset were good and stable, whether some variables were linear combinations of others or not. Moreover, when τ and λ were chosen independently, the proposed method proved to be robust to the choices of these hyper-parameters: only 1 run among 24 in the sensitivity analysis gave bad results.

In classical cases using the g-prior, many authors suggested to put prior distributions on τ, see Section 1. Following them, an idea would be to put prior distributions on the hyper-parameters τ and λ. However, these authors often used Bayes Factors and not a latent γ vector like us. They were then more in the spirit of model selection than in the spirit of variable selection. The choice of τ can have influence on the posterior probabilities of models, and therefore on Bayes Factors (see Celeux et al. [2006] for instance). On the opposite, from our experience, methods of variable selection using a latent γ vector are not overly sensitive to the value of τ. Considering the facts that the variable selection algorithm proposed in this paper uses a latent γ vector and that this algorithm appeared robust to the values of τ and λ, we did not put prior distributions on these hyper-parameters. All the more that putting prior distributions on these hyper-parameters would lead to a non-standard posterior distribution for τ. In practice, even if $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ is theoretically invertible, some variables can be highly correlated and $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ can be computationally singular. Moreover, we do not necessarily know if some variables are linear combinations of others. To avoid a computational problem using the classical g-prior, we suggest to use the prior and the algorithm proposed in this paper, even if eventually $\mathbf{X}_\gamma^T\mathbf{X}_\gamma$ is never singular. Once a final selection of variables $\gamma+$ is obtained by our algorithm, the rank of the matrix with all the variables finally selected, denoted by $X_{\gamma+}$, should be computed. If this matrix is not of full rank, some variables are linear combinations of others, and we can take a submatrix of $X_{\gamma+}$ of full rank as a new data matrix. Note that it is easier to take linearly independent columns of $X_{\gamma+}$, than linearly independent columns of X, especially if p is quite large.

The result of a run of the proposed algorithm is a vector giving the number of iterations during which variables have been selected. We suggested to represent this vector by a boxplot to decide which variables should be in the final selection, by taking the variables which distinguished from others using a threshold. However, it would be interesting to have a non-supervised criteria to decide which variables should be in the final selection of a run. Finally, a direction for future research is to use the proposed prior for β_γ in the framework of variable selection for generalized linear mixed models. Gupta and Ibrahim [2009] proposed an Information Matrix Ridge prior for generalized linear models, but not in an objective of variable selection.

References

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

M.C. Baragatti. Bayesian variable selection for probit mixed models applied to gene selection. *arXiv:1101.4577*, 2011.

M.C. Baragatti and D. Pommeret. Comment on "bayesian variable selection for disease classification using gene expression data". *Bioinformatics*, 2011.

L. Bottolo and S. Richardson. Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

P.J. Brown, M. Vannucci, and T. Fearn. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society B*, 60(3):627–641, 1998.

P.J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society B*, 64(3):519–536, 2002.

G. Celeux, J.M. Marin, and C.P. Robert. Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147:59–79, 2006.

M.H. Chen and J.G. Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, 13:461–476, 2003.

H. Chipman, E.I. George, and R.E. McCulloch. The practical implementation of bayesian model selection. In *Model selection - IMS Lecture Notes*. P. LAHIRI. Institute of Mathematical Statistics, 2001.

W. Cui and E.I. George. Empirical Bayes vs. fully Bayes variable selection. *J. Stat. Plann. Inference*, 138(4):888–900, 2008. doi: 10.1016/j.jspi.2007.02.011.

E.I. George and D.P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87 (4):731–747, 2000. doi: 10.1093/biomet/87.4.731.

E.I. George and R.E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

M. Gupta and J.G. Ibrahim. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 102(479):867–880, 2007. doi: 10.1198/016214507000000068.

M. Gupta and J.G. Ibrahim. An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, 19(4):1641–1663, 2009.

K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, and B.K. Mallick. Gene selection: a bayesian variable selection approach. *Bioinformatics*, 19(1):90–97, 2003.

F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. doi: 10.1198/016214507000001337.

J.S. Liu. The collapsed gibbs sampler in bayesian computations with application to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

J.M. Marin and C.P. Robert. *Bayesian core: a practical approach to computational Bayesian statistics.* Springer-verlag inc edition, 2007.

D.W. Marquardt. Generalized inverses, ridge regression, biaised linear estimation, and nonlinear estimation. *Technometrics*, 3:591–612, 1970.

Y. Maruyama and E.I. George. A g-prior extension for p¿n. *arxiv:0801.4410*, 2010.

G.O. Roberts and J.S. Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability*, 16(4):2123–2139, 2006.

M. Smith and R. Kohn. Non parametric regression using bayesian variable selection. *Journal of Econometrics*, 75:317–344, 1997.

P. Somol and J. Novovicova. Evaluating the stability of feature selectors that optimize feature subset cardinality. In N. da Vitora Lobo et al., editor, *Lecture Notes in Computer Science, vol 5342*, pages 956–966. Springer-Verlag Berlin Heidelberg, 2008.

D.A. van Dyk and T. Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103:790–796, 2008.

M. West. *Bayesian Statistics 7*, chapter Bayesian factor regression models in the 'Large p, Small n'paradigm. 2003.

A.J. Yang and X.Y Song. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2):215–222, 2010.

A. Zellner. *Bayesian Inference and Decision Techniques – Essays in honour of Bruno De Finetti.*, chapter On assessing prior distributions and Bayesian regression analysis with g-prior distributions., pages 233–243. Amsterdam, 1986.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603. University of Valencia Press, 1980.